AUTOMATED ANALYSIS OF RADAR IMAGERY OF VENUS: HANDLING LACK OF GROUND TRUTH

M. C. Burl[†‡], *Usama M. Fayyad*[†], *Pietro Perona*[‡], *and Padhraic Smyth*[†]

[†] Jet Propulsion Laboratory
California Institute of Technology
MS 525-3660 — Pasadena, CA 91109
{fayyad,pjs}@aig.jpl.nasa.gov

[‡] Electrical Engineering
California Institute of Technology
MS 116-81- Pasadena, CA 91125
{burl,perona}@systems.caltech.edu

ABSTRACT

Lack of verifiable ground truth is a common problem in remote sensing image analysis. For example, consider the synthetic aperture radar (SAR) image data of Venus obtained by the Magellan spacecraft, Planetary scientists are interested in automatically cataloging the locations of all the small volcanoes in this data set; however, the problem is very difficult and cannot be performed with perfect reliability even by human experts. Thus, training and evaluating the performance of an automatic algorithm on this data set must be handled carefully. We discuss the use of weighted free-response receiver-operating characteristics (wFROCs) for evaluating detection performance when the "ground truth" is subjective. in particular, we evaluate the relative detection performance of humans and automatic algorithms. Our experimental results indicate that proper assessment of the uncertainty in "ground truth" is essential in applications of this nature.

1. INTRODUCTION

Very large image databases are becoming more prevalent in a variety of scientific, medical, and engineering disciplines. In planetary science and astronomy, hardware advances in recent years have led to increases of several orders of magnitude in the volume of data returned per instrument or per spacecraft. In particular, the Magellan spacecraft which recently surveyed the planet Venus returned more data to Earth than all previous inter-planetary missions combined.

In planetary science, image analysis has t raditionally been, and often still is, a strictly manual process. For example, feature catalogs are typically constructed by careful visual inspection, i dent ification, and measurement of geologic features in a set of hardcopy photographs. However, due to the increased size of image databases currently being collected, simple manual cataloging is no longer a practical option -- especially if any significant fraction of the total available data is to be utilized.

We have been involved in the development and deployment of tools for the automated analysis of large astronomy and planetary image databases [I]. This paper focuses on the image data set obtained by the Magellan spacecraft at Venus. We are ultimately targeting the development of a t rainable image analysis system with built-in learning components. A scientist *trains* the system to find objects of interest by simply giving it examples of the target objects. In addition to automating laborious and visually-intensive tasks, such a system provides an objective, and repeatable process, thereby enabling the scientists to base their analyses on uniformly consistent data, with subjective variations minimized.

This paper focuses in particular on issues that arise when the training data cannot be considered '(ground truth" in the usual sense, i.e., when objects are located by the scientists subjectively. This effectively introduces noise into the training and evaluation of the detection system. Also, it naturally raises the questions of how accurate any one human detector may be, how humans compare with each other, and how well an algorithm may be expected to perform relative to human performance.

Previous work has dealt with some of the general theoretical aspects of "noisy class labels for supervised pattern recognition [2, 3]. In addition, there has been considerable work in the statistical literature on the topic of combining multiple subjective estimates [4, 5]. The originality of the work described in this paper lies in handling the ground truth ambiguity problem in the context of a large-scale, real-world, image analysis problem. Subjective evaluations are commonly used in a variety of

other image analysis applications (in remote sensing and medical diagnosis) where verifiable ground truth is impractical to obtain due to the associated costs or risks.

2. AUTOMATED VOLCANO CATALOGING IN SAR IMAGES

The Magellan spacecraft transmitted back to earth a data set consisting of over 30,000 high resolution synthetic aperture radar (SAR) images of the Venusian surface. It was necessary to use radar in order to penetrate the opaque cloud cover surrounding Venus. The primary radar imaging parameters were as follows: the SAR frequency was 2.385 GHz, the full-resolution pixel-spacing was 75m, the radar incidence angle ranged from 15 deg to 45 deg, and the number of "radar looks" varied from 5 to 16 [6]. Each radar image is 1024 pixels square. The data is publically available at minimal cost from NASA as a set of about 100 CD-ROMS [7].

This is by far the most detailed data set ever assembled for any of the planets. The data represents a treasure-trove of potential scientific information for planetary scientists. The study of volcanic processes is essential to an understanding of the geologic evolution of Venus. Central to volcanic studies is cataloging the location, size, and characteristics of each volcano. However, there are estimated to be on the order of 10^6 visible volcanoes scattered throughout the 30,000 Magellan images [8]. Manually locating all of these volcanoes would require on the order of 10 man-years of a planetary geologist's time, notwithstanding the disadvantages of deriving a catalog in this manner (subjective bias, non-repeatable).

We have previously presented empirical detection performance results for an automatic algorithm based on spatial eigenrepresentations and supervised classification [9]. The algorithm's performance was shown to be comparable to that of planetary scientists. The pattern recognition system which forms the basis of these results uses a matched filter (for example, the mean of locally windowed training examples of volcanoes) to initially focus attention on local regions of interest. The detected local regions are then projected into a subspace consisting of significant principal directions of the training data. This subspace is determined by selecting the most significant components produced by a singular value decomposition (SVD) of the training data. The SVD approach has been used elsewhere for recognition problems [1 O] and has some well-known weaknesses, such as sensitivity to scale and translation. For the volcano application, there is relatively little scale and rotation variation. The focus of attention (FOA) step is quite effective at accurately locating the centers of the volcanoes, thus minimizing any translation effects. The significant SVD responses are fed to a classifier trained to discriminate between volcano and non-volcano local regions resulting from the FOA stage. Classification in the projected subspace using a simple maximum-likelihood Gaussian classifier with full covariance matrices has been found to perform as well as alternative non-parametric methods such as neural networks and decision trees [9].

The system is trained as follows. First, a FOA filter is constructed from the set of all training volcanoes (windowed to a fixed size, say 30 x 30 pixels). Second, the SVD basis is determined from the same set of training volcanoes. Finally, a filtered and projected set of "candidate" local regions are seperated into "volcano" and "non-volcano" regions -- by matching them with a reference list of volcanoes. Each step relies on the availability of a set of images within which the volcanoes have been labelled. "Labelling" consists of having a scientist examine an image and produce a list of x,y coordinates describing where the volcanoes exist in the image (if any). An interactive graphical interface has been developed to assist in this process, allowing the scientist to simply point and click on targets in training images.

3. AMBIGUITY IN VOLCANO DETECTION

3.1, Volcano Identification is Subjective

Figure 1 shows a typical sub-image from the Magellan set with a number of small volcanoes present. The radar illumination is from the left, and thus the larger volcanoes display a readily-visible characteristic appearance: a bright upward-sloping left flank and a dark downward sloping right flank. Many volcanoes also have a "dark-bright" pattern at the center caused by the presence of a summit pit. These visual cues are the primary ones used by the planetary scientists to locate volcanoes. Some volcanoes do not have the characteristic appearance, but instead can be identified based on local changes in texture, radial flow patterns, or disruption of surrounding linear features.

Identifying volcanoes in the Magellan images is non-trivial for a number of reasons. The volcanoes themselves exhibit considerable variety, both in terms of underlying topography and shape, and

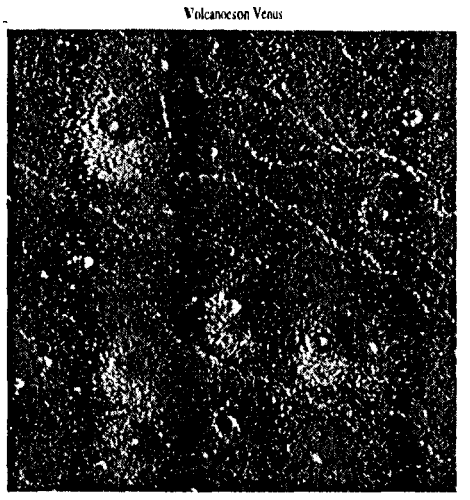

Figure 1: Magellan SAR sub-image: A 30km^2 area containing a number of small volcanoes, left illumination, inc. angle $\approx$ 40°.

in the manner in which they respond to the radar illumination.]n addition, the volcanoes appear in a variety of geologic contexts, often with visually confounding backgrounds, e.g., volcanoes can occur superposed on other geologic features such as dense collections of ridges,

The relatively low signal-to-noise ratio of the volcanoes over the background can cause considerable labelling variability even when experienced scientists label the same image. For example, we have found with our standard training set of 4 images that typically, for any pair of scientists, about 70-80% of these identifications will be common to both scientists, and 30-40% will be unique to each, '1'bus, if onc scientist were arbitrarily designated "ground truth", the the other scientist would typically have a 70-80% detection rate and a 30-40% false alarm rate,

3.2. The Use of Rating Categories

In order to better model the subjective uncertainty present in the labelling process, the labeler not only identifies the location of each volcano, but also provides a subjective assessment of his certainty that a volcano exists at that location. It is well known that direct elicitation of subjective probabilities from humans is quite difficult and prone to various calibration errors and biases [11]. A more effective approach in practice is to have the scientists label training examples into quantized probability bins, where the probability bins correspond to visually distinguishable sub-categories of volcanoes. In particular, wc have used five categories: (i) volcanoes having clearly visible summit pits, bright-dark radar pair, and apparent topographic s1OPC, probability 0,98, (ii) only 2 of the 3 criteria in the first category are visible, probability 0.80, (iii) no summit pit visible, however, there is evidence of flanks or circular outline, probability 0.60, (iv) only a summit pit visible, probability 0.50, (v) no volcaxlo-like features visible, probability 0.0. The probabilities correspond to the mean probability that an object is indeed a volcano given that it has received a particular category label. These mean values were elicited after considerable discussions with the participating planetary geologists concerning their validity and interpretation. Wc refer to labels categorized in this manner as "categorized probability labels."

Figure 2 shows some typical volcanoes from each category. While this simple quantized method may not fully capture the uncertainty of the labeller it is certainly a much more useful approach than forcing the labeller to make a binary class decision (as wc shall see later). The use of quantized probability bins to attach levels of certainty to subjective image labelling is not new: the same approach is routinely used in the evaluation of radiographic image displays to generate subjective receiver operating characteristics (ROCs) [12]. However, in this paper wc extend the basic approach by defining the notion of weighted ROCs (Section 4).

4. PERFORMANCE EVALUATIOA^T METHODOLOGIES

4.1. The Free-Response ROC Method

Standard ROC methodology plots the probability of detection as a function of the probability of false alarm --- the ROC is an implicit function of a threshold which can be applied to the the decision-making system's output, For the problem of object detection in images, the probability of false alarms is not well defined; instead, probability of detection is plotted as a function of the false alarm *rate* relative to the total estimated number of volcanoes: this results in a slightly modified ROC methodology, equivalent to the free-response ROC (FROC) [13, 14].

Since we do not have a reference ground truth list for the volcano detection problem, how should detection performance be measured? Onc approach is to use consensus-based estimates as the reference ground-truth, i.e., have a consensus of scientists label the images, treat these as ground truth, and

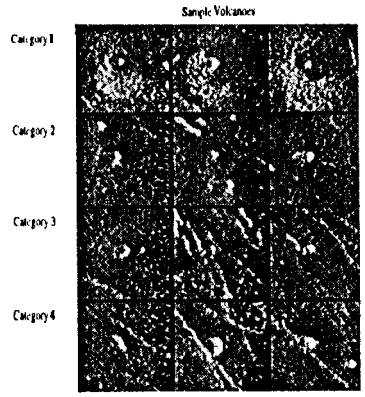

Figure 2: A small selection of volcanoes from four categories as labeled by the geologists.

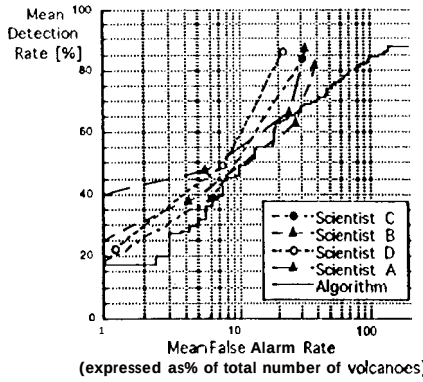

Figure 3: FROC Plot Showing Individual Scientists Vs. Consensus Labelling

evaluate the FROC performance of the other scientists relative to this consensus "truth ." The use of categorized probability labels with k bins allows the definition of k points on an FROC curve: the first point comes from the detection/false-alarn~ performance resulting from only using category 1's as volcanoes, while treating all other categories as false alarms; the second point admits only category 1 and 2 volcanoes; and so on.

Figure 3 shows such an FROC plot based on 4 test images containing an estimated 150 volcanoes. Four individual scientists were compared with the consensus generated by two of them (A and B). Although the consensus labelling was generated some time after the individual labelling, there is undoubtedly artifactual correlation between the labellings of A and B and their consensus. The performance of the detection algorithm described in Section 2 is also shown. Here, the FROC curve is the cumulative performance when the algorithm was trained in cross-validation mode on each set of 3 images and evaluated on the fourth. Note that the algorithm appears comparable with the performance of the scientists at the lower detection/false-alarnl rate regions, but is less competitive at the higher end.

4,2. The Weighted FROC Method

As described above, the standard FROC approach assumes that ground truth is known. However, rather than using the consensus list as absolute ground truth, one can instead treat the consensus list probabilistically. In this case, each local region in the consensus is considered to be a detection with probability p and a false alarm with probability $1 - p$, where the probability p is determined by the category label (confidence) that was given to the region by the consensus labellers. This results in a weighted FROC (wFROC) measure of performance. The overall effect is to drag the standard FROC (where no allowance is made for the probabilistic effect) towards the "center" of the plot, away from the ideal "false alarm rate 0.0, detection rate 1.0" operating point. Furthermore, the ideal "perfect" operating point is no longer achievable by *any* system, since the reference data is itself probabilistic. Hence, an effective optimal wFROC is defined by exactly matching the probabilistic predictions of the reference list — one can do no better.

Figure 4 shows the same data as plotted in Figure 3 but now evaluated as a wFROC instead of an FROC. We note two primary effects relative to the standard FROC:

- For any fixed false alarm rate, the detection rates are now less optimistic than with the FROC, for both humans and algorithm.
- The algorithm's curve has separated from the humans in the wFROC. The SVD algorithm appears to be doing poorly in terms of approximating posterior probabilities and wFROC is more sensitive to this than the FROC. This poor performance appears to be due to the fact that the subspace projection destroys the implicit probabilistic information present in the categories. Category 1 's, 2's, etc. are jumbled up in the projected space without any obvious structure.

5. DISCUSSION AND CONCLUSION

The wFROC method is a useful step in the direction of quantifying subjective uncertain y in labelled

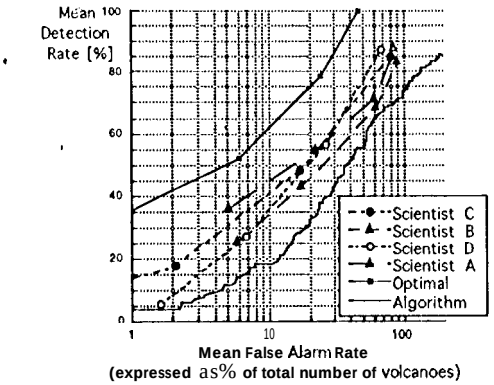

Figure 4: FROC Plot Showing Individual Scientists Vs. Consensus Labelling

training data for image analysis. Currently we are investigating statistical estimation techniques which combine individual labellings into a composite probability estimate for each local region [15, 16]. In this manner a given labeller can be compared with the mathematical consensus of all the' other labellers. The quantitative details differ from the simpler method presented above, but the qualitative results are the same: (a) ignoring label uncertainty can lead to over-optimistic estimates of both human and algorithmic performance, and (b) proper treatment of label uncertainty can reveal differences between human and algorithmic performance which may be hidden by simpler methods.

The proposed wFROC technique provides a framework for more accurate estimation and evaluation of basic image quantities of interest for applications where absolute ground truth is not available. Such applications are becoming increasingly common as remote-sensing platforms provide orders of magnitude more data and well-calibrated ground truth constitutes a tiny (and perhaps even zero) fraction of the overall data set.

Acknowledgements

The authors would like to thank Jayne Aubele and Larry Crumpler of the Department of Geological Sciences, Brown (University, for their assistance in labelling images, and Maureen Burl (JPL) for assistance with the experiments. The research described in this paper was carried out by the Jet Propulsion Laboratory, California institute of Technology, under a contract with the National Aeronautics and Space Administration.

6. REFERENCES

[1] U. M. Fayyad, P. Smyth, N. Weir, and S. Djorgovski, "Automated analysis and exploration of large image databases: results, progress, and challenges," *Journal of Intelligent Information Systems*, in press.

[2] G. Lugosi, "Learning with an unreliable teacher," *Pattern Recognition*, vol. 25, no.1, pp.79- 87, 1992.

[3] B. Silverman, "Some asymptotic properties of the probabilistic teacher," *IEEE Trans. Info. Theory*, 11'-26, I1o.2, pp.246 249, 1980.

[4] J. S. Uebersax, "Statistical modeling of expert ratings on medical treatment appropriate ness," *J. Amer. Statist. Assoc.*, vol.88, no.422, pp.421–427, 1993.

[q A. Agresti, "Modelling patterns of agreement and disagreement," Statistical *Methods in Medical Research*, vol.1, pp.201- *218, 1992.*

[6] J. P. Ford et al, *Spaceborne Radar* Observations: *A Guide for Magellan Radar Image Analysis*, Jet l'repulsion Laboratory Publication 89-41, JPL, Pasadena, CA, 1989.

[7] *NSSDC News*, VOI.1O, no.], Spring *1994*, available from *request @nssdc.gsfc.nasa.gov.*

[8] J. C. Aubelle and E. N. Slyuta, "Small domes cm Venus: characteristics and origins," in *Earth, Moon and Planets*, 50/51, 493-532, 1990.

[9] M. C. Burl, U. M., Fayyad, P. Perona, P. Smyth, and M. P. Burl, "Automating the hunt for volcaUocs on Venus," in Proceedings *of the* 1994 *Computer Vision and* Pattern Recognition Conference, *CVPR-94*, Los Alamitos, CA: IEEE Computer Society Press, pp.302-309, 1994.

[10] M. Turk and A. Pentland. "Eigenfaces for recognition." *J. of* Cognitive *Neurosci.*, 3:71-86, 1991.

[1] I). Kahneman, P. Slovic, and A. Tversky (eds.), *Judgement under [Uncertainty: Heuristics and Biases*, Cambridge University Press, 1982.

[2] M. S. Chesters, "Human visual perception and ROC methodology in medical imaging," *Phys. Med. Biol.*, vol.37, no.7, pp.1433-1476, 1992.

[]3] P. C. Bunch, J. F. Hamilton, G. K. Sanderson and A. H. Simmons, "A Free-Response approach to the measurement and characterization of radiographic-observer perform ance," *J. Appl. Photo. Eng.*, vol.4, 110.4, pp.166-171, 1978.

[14] I). P. Chakraborty and 1,. H. I,. Winter, "Free-Response methodology: alternate analysis and a new observer-performance experiment," *Radiology, 174, 873-881, 1990.*

[15] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied Statistics*, vol.28, no.1, pp.20-28, 1979.

[16] P. Smyth, U. Fayyad, M. Burl, 1'. Baldi, P. Perona, 'Inferring ground truth from subjective labelling of radar images of Venus,' in preparation.